IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steven Han
December 30, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
- Data Collection via API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) using SQL, Pandas and Matplotlib
- Interactive Visual Analytics
    - Created map of launch sites and success rates using Folium
    - Built an interactive Dashboard using Plotly Dash
- Machine Learning Prediction Models
    - Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbors (KNN)
    - Tuned hyperparameters with GridSearchCV
- Summary of all results
- EDA results
- Interactive analytics
- Predictive analysis

Introduction

- Project background and context
  - SpaceX offers Falcon 9 launches at ~ $62M, far below competitors ($165M+)
  - Reusable first-stage boosters are the primary driver of this cost advantage
  - Booster recovery success has a direct impact on launch pricing and profitability
  - Predicting landing outcomes enables more accurate cost estimates
- Problems you want to find answers
  - How can we use machine learning models to predict Falcon 9 booster landing success before launch?
  - What factors influence the success rate of landing the first-stage booster?
  - Which machine learning model provides the best prediction of a successful landing?

Section 1

Methodology

Methodology

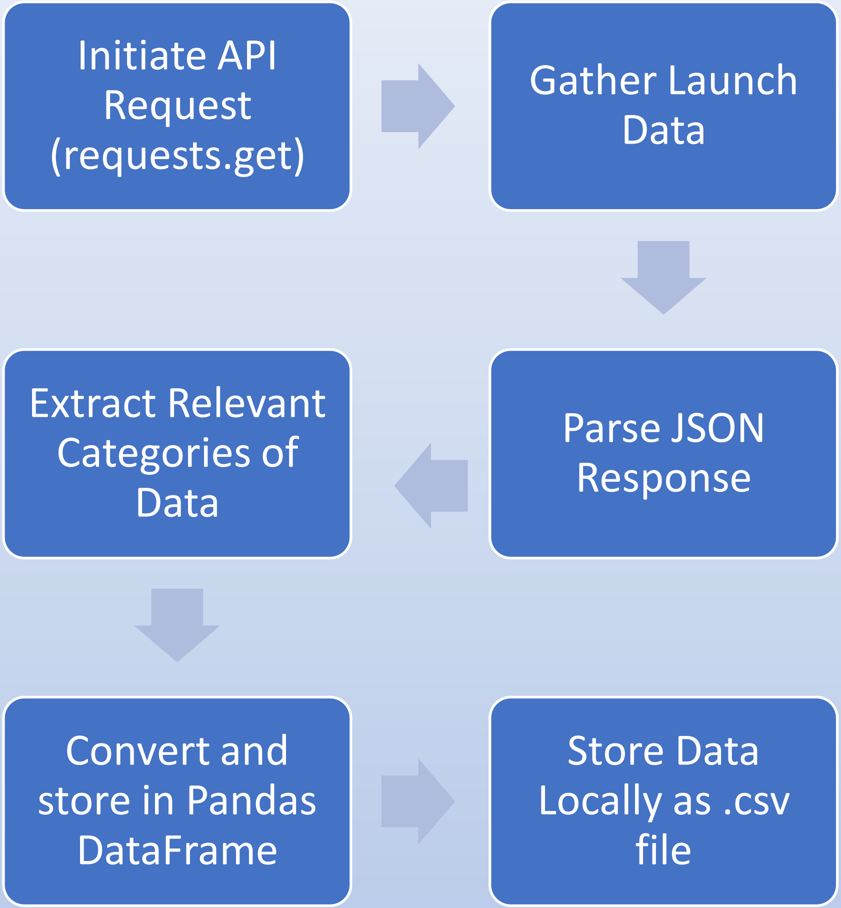
Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX REST API and Web Scraping from Wikipedia page “List of Falcon 9 and Falcon Heavy Launches”
- Perform data wrangling
  - SQL was used to query, filter and structure the raw data
  - Pandas was used for data cleaning, transformation, standardization and dealing with missing values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification models were tested – including Logistic Regression, KNN, SVM, and Decision Trees – to determine the best-performing model

Data Collection – SpaceX API

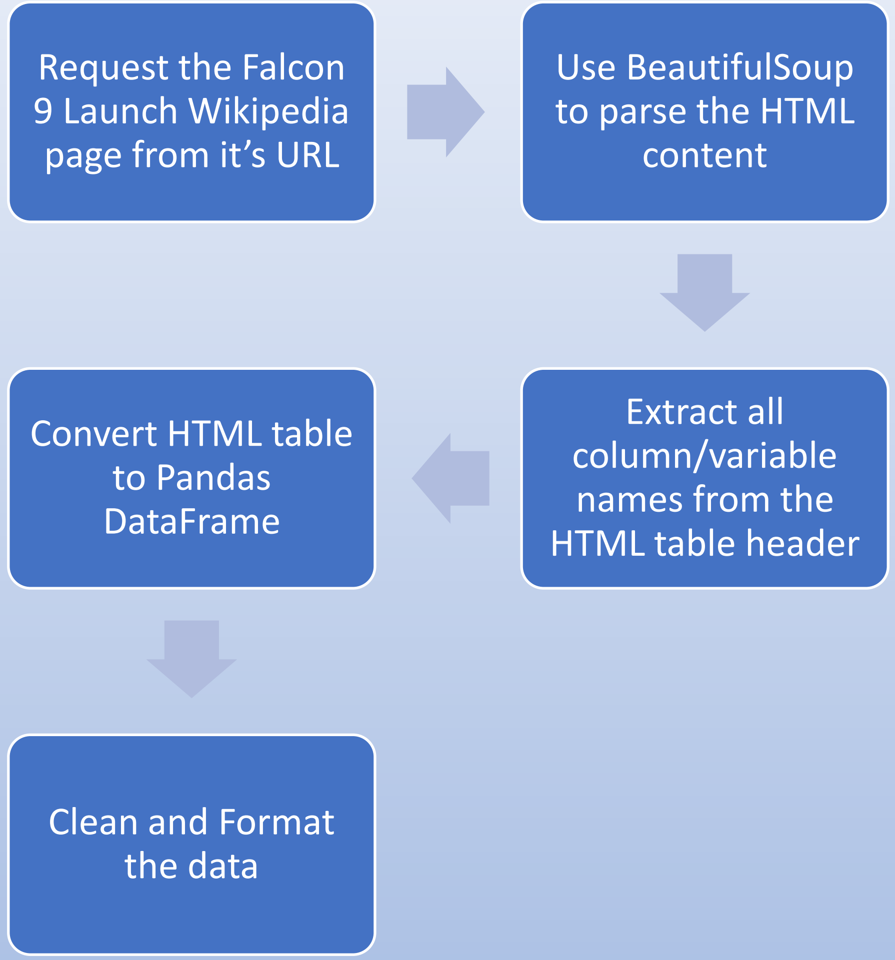
- Data collection with SpaceX REST calls using key phrases and flowcharts

- GitHub URL of the completed SpaceX API calls notebook [SpaceX API notebook hyperlink]



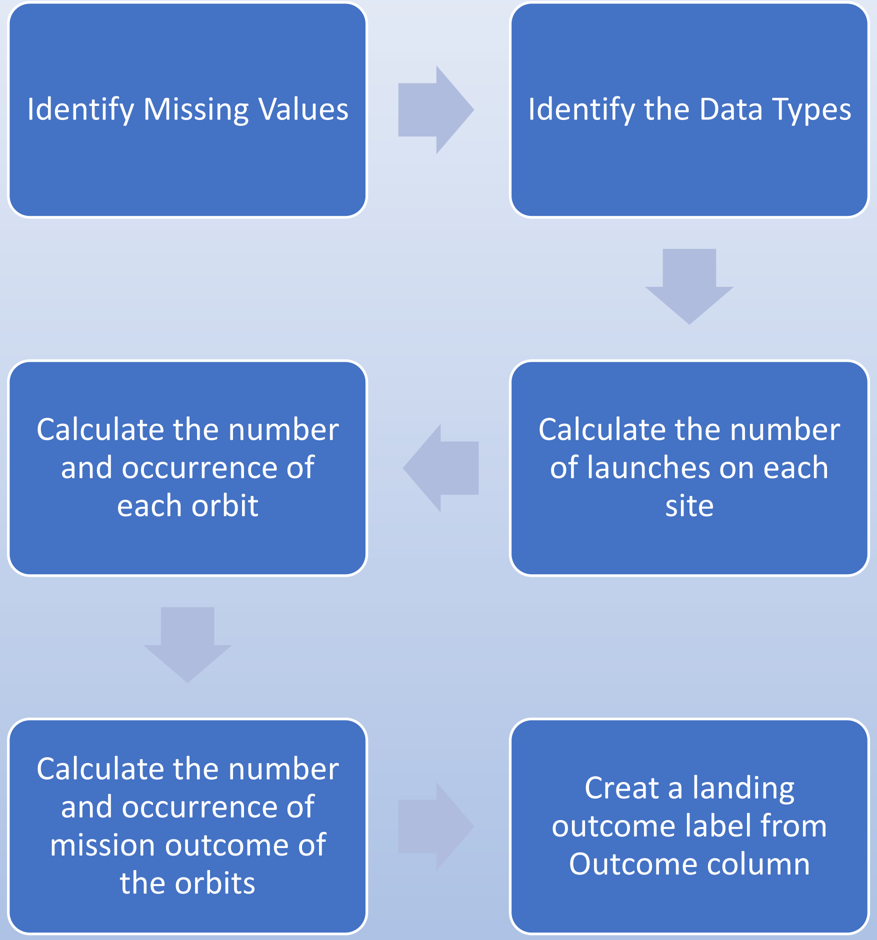
Data Collection - Scraping

- Web Scraping process using key phrases and flowcharts
- GitHub URL of the completed web scraping notebook Web Scraping notebook hyperlink



Data Wrangling

- Data wrangling process using key phrases and flowcharts
- GitHub URL of the completed data wrangling related notebooks, [Data Wrangling notebook hyperlink]()



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Scatterplot to visualize the relationship between Flight Number and Payload Mass
  - Scatterplot to visualize the relationship between Flight Number and Launch Site
  - Scatterplot to visualize the relationship between Payload Mass and Lauch Site
  - Bar chart to visualize the Success Rate by Orbit Type
  - Scatterplot to visualize the relationship between Flight Number and Orbit Type
  - Scatterplot to visualize the relationship between Payload Mass and Orbit Type
  - Line chart showing Launch Success Rate Trend by Year

- GitHub URL of the completed EDA with data visualization notebook, [EDA with Data Visualization notebook hyperlink]

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string ‘CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS):
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List all the booster versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function
  - List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site and the months in year 2015
  - Rank the count of landing outcomes (such as Failure(drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL of the completed EDA with SQL notebook, EDA with SQL notebook hyperlink

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Markers were added to mark the launch sites on the map with longitude and latitude coordinates. This makes launch site easily identifiable.
  - Circles show the area around the launch sites. Added to show the general vicinity of the launch sites.
  - Marker Clusters added to group many nearby points into a single icon to prevent map clutter.
  - Lines were added to illustrates distances from the launch site to points, such as, nearest cities, highways and railroads.

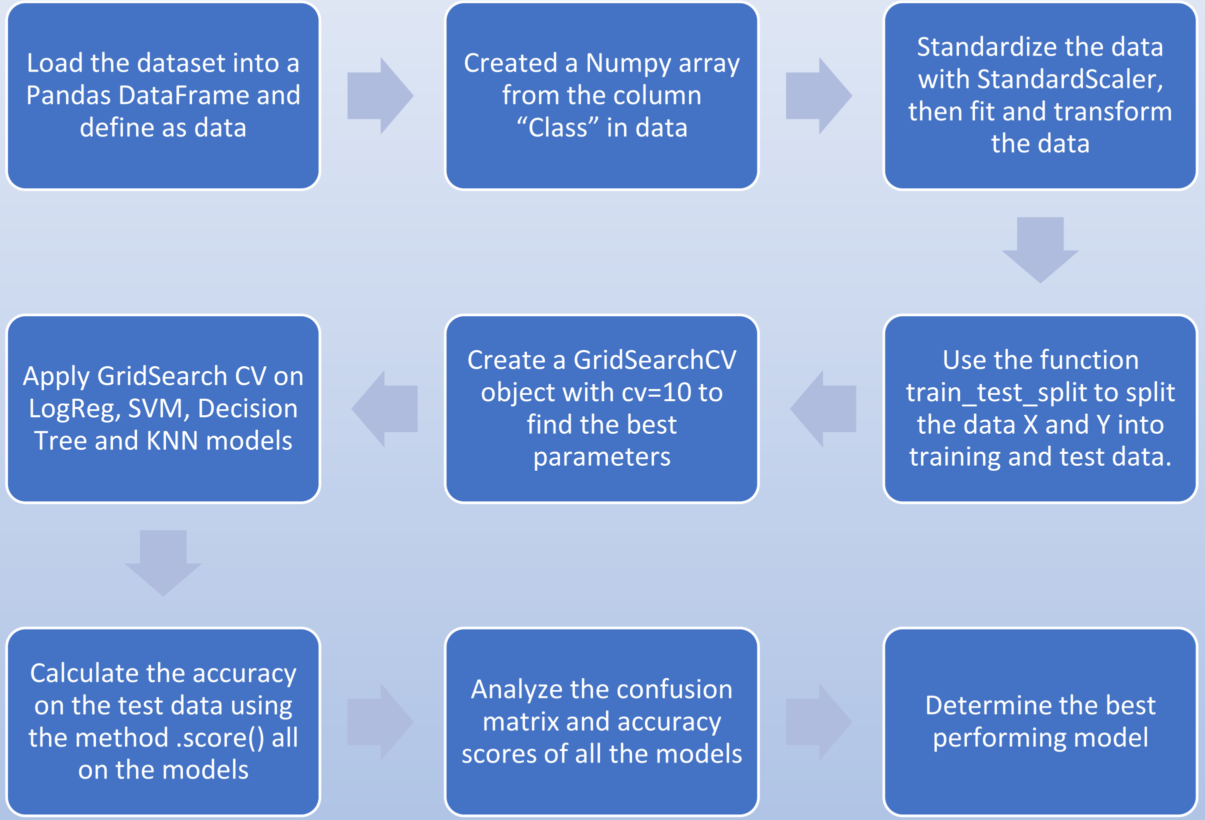
GitHub URL of the completed interactive map with Folium map, [Interactive Map with Folium notebook hyperlink]

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Interactive Pie Chart with dropdown menu of Total Successful Launches by All Sites or Unique Sites
  - Scatter Plot of Payload vs. Launch Outcome for All Sites with interactive Payload Range Slider.
- Pie chart helps visualize the overall success rate and the success/fail rate on individual launch sites.
- Scatter Plot with Range Slider allows users to explore how payload mass affects mission outcomes.
- GitHub URL of the completed Plotly Dash lab, Dashboard with Plotly Dash notebook hyperlink

Predictive Analysis (Classification)

- Summary of the model development process after loading the dataset.
  - Created a NumPy array from the column“Class” in data
  - Standardized the data
  - Split the data to train and test sets
  - Apply GridSearch CV on different models: Logistic Regression, SVM, Decision Tree, KNN
  - Review results and determine best performing model
- GitHub URL of the completed predictive analysis lab, [Predictive Analysis notebook hyperlink]



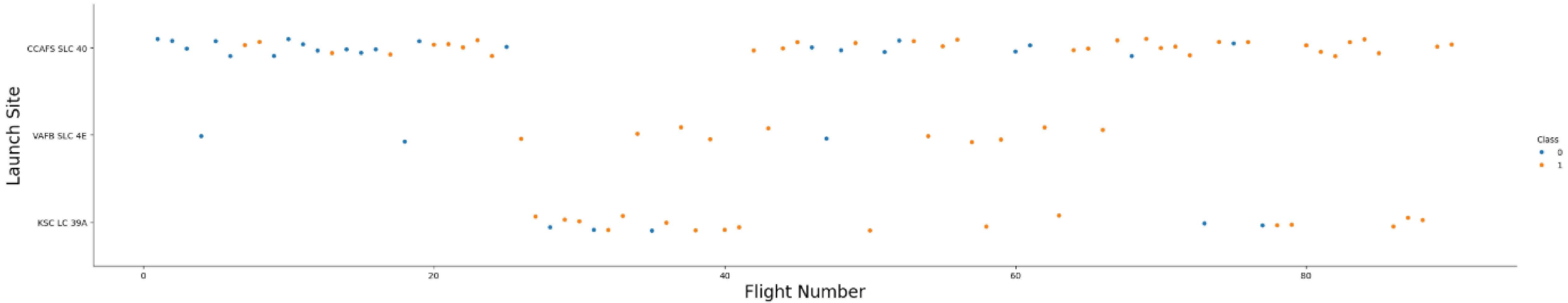
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

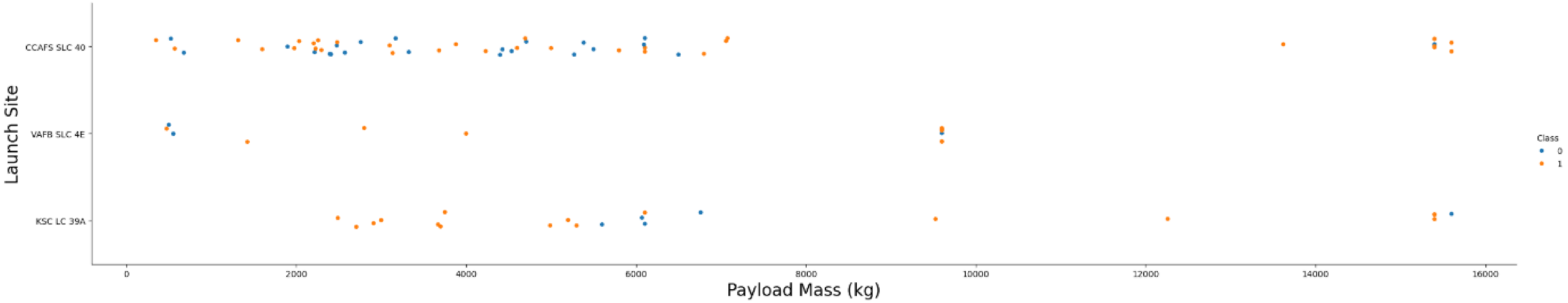
Insights drawn from EDA

Flight Number vs. Launch Site



- Observed Patterns show:
  - More flights and experience increase successful landings across all launch sites.
  - CCAFS SLC-40 has the highest launch frequency and shows a clear transition from mixed outcomes to mostly successful landings in later flights.
  - Overall, the plot suggests that experience and operational maturity play a significant role in landing success.

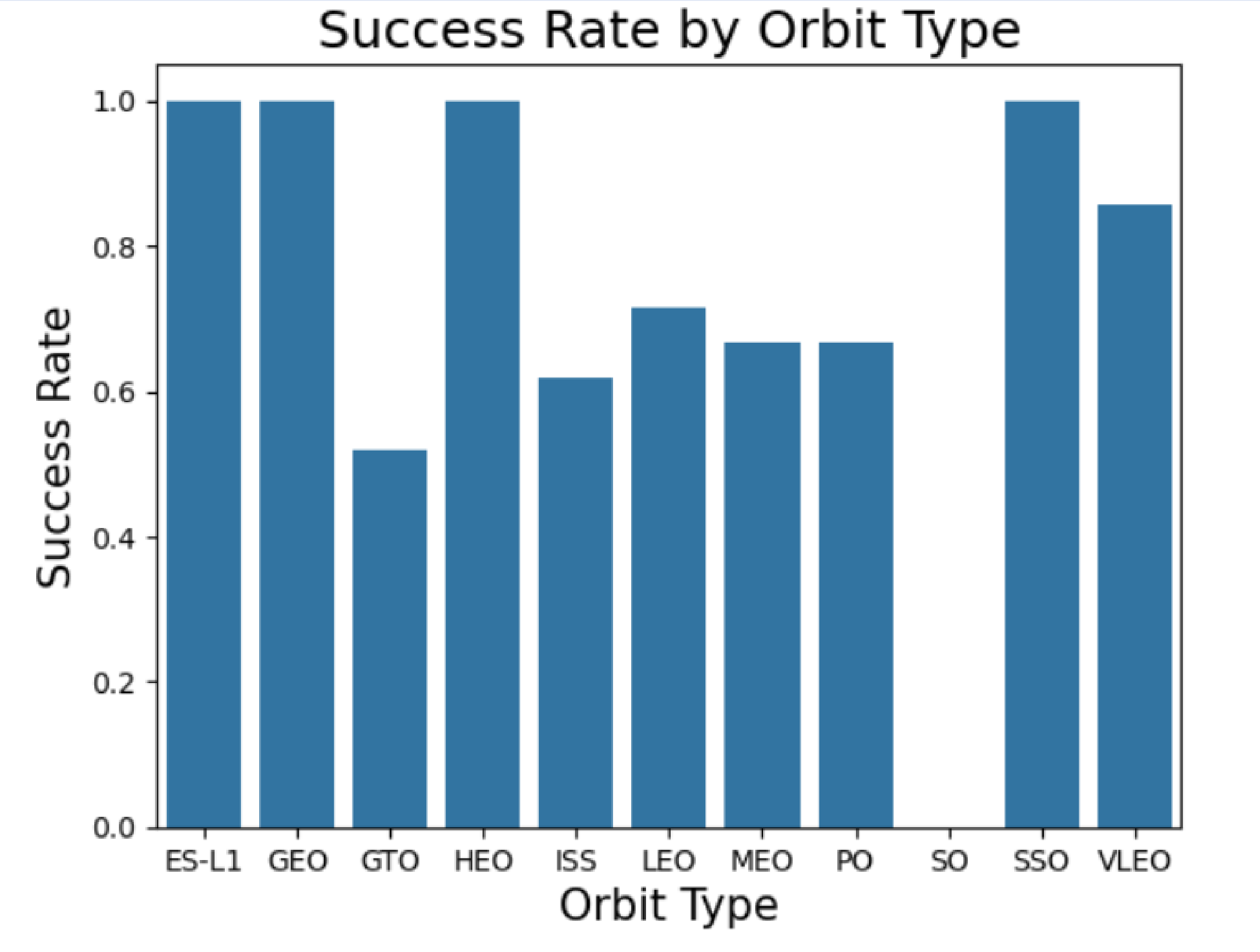
Payload vs. Launch Site



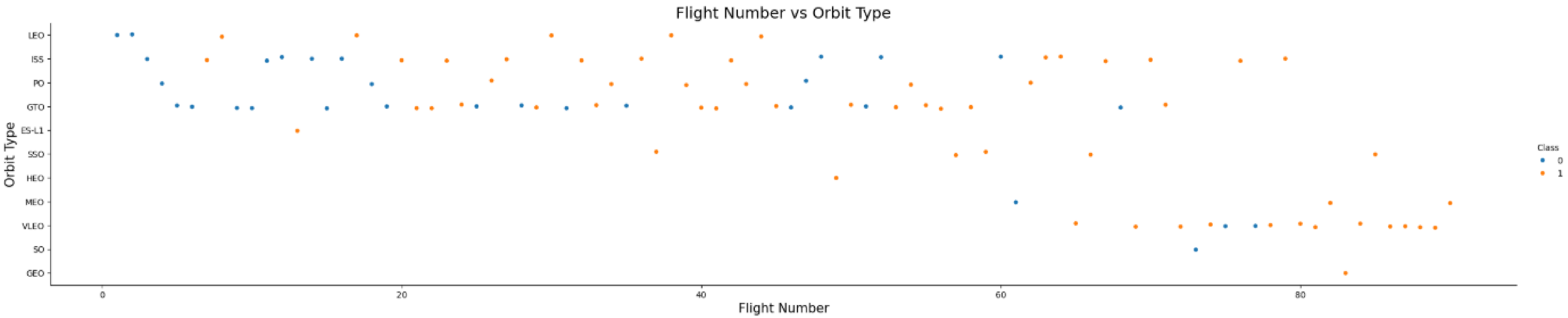
- While there are less launches with Payload Mass over 8,000 kg, the success rate is greater. Note that for VAFB-SLC launch site, there are no rockets launched for Payload Mass over 10,000 kg.
- CCAFS SLC 40 is the most active launch site.

Success Rate vs. Orbit Type

- Several Orbit Types have a 100% success rate: ES-L1, GEO, HEO, SSO, indicting high confidence for a successful mission.
- Orbit Type SO has a 0% success rate, indicating more research and innovation is need for a successful mission.
- Orbit Types GTO, ISS, LEO, MEO, PO and VLEO are fairly reliable with all achieving over a 50% success rate.

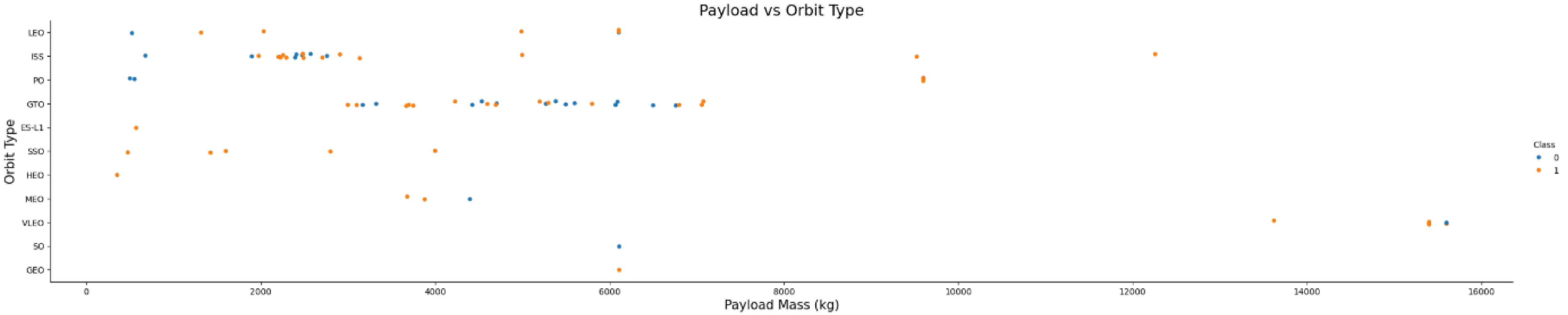


Flight Number vs. Orbit Type



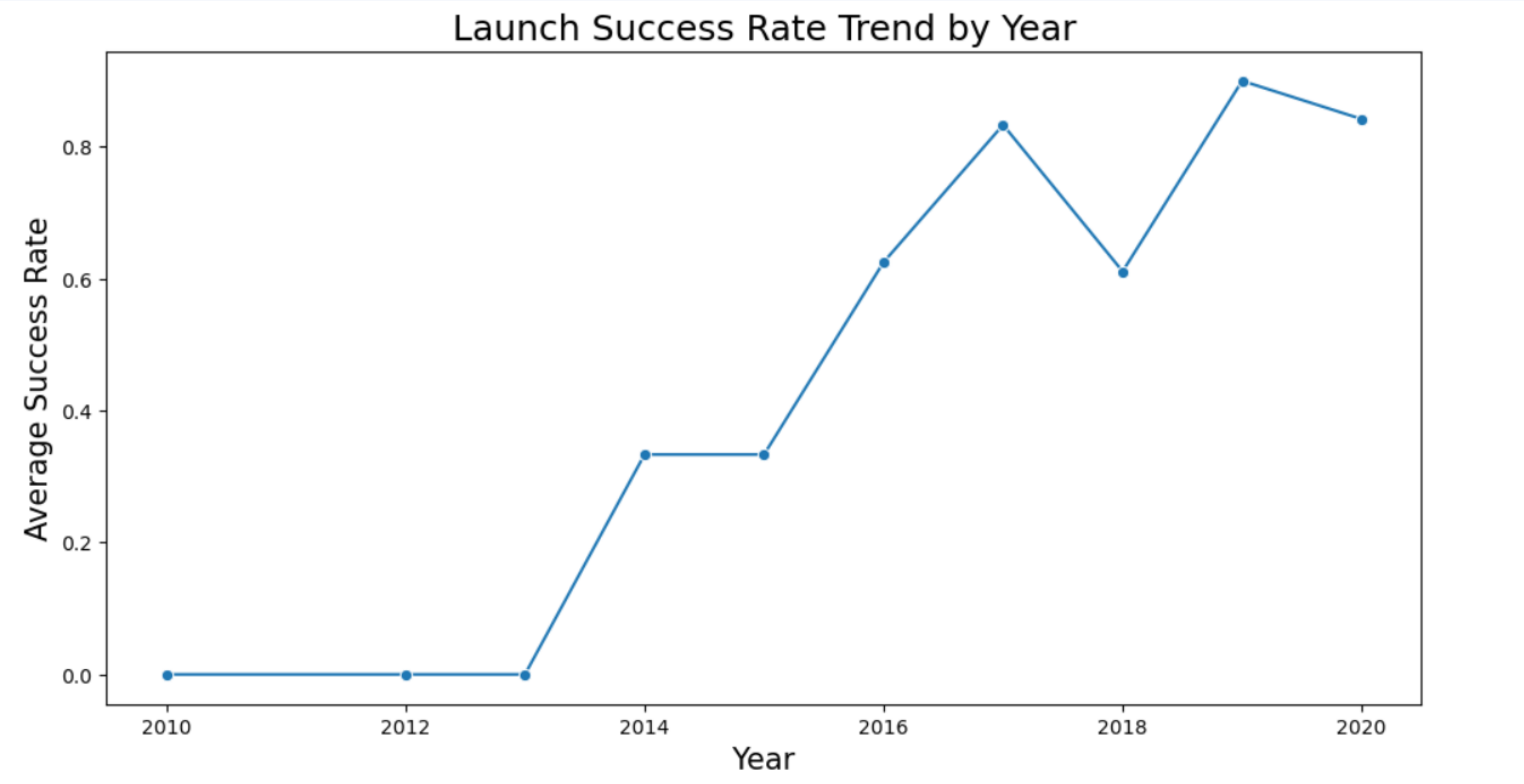
- The LEO orbit success rate appears related to the number of flights. While on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both successful and unsuccessful landing rates are prevalent.

Launch Success Yearly Trend



- Since 2013 the success rate has steadily increased other than slight decreases in 2018 and 2020.

All Launch Site Names

- Query in screenshot, the result displays the unique launch site names: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

```
%%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

[18]:
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Query in screenshot, the result displays 5 records where launch sites begin with ‘CCA’.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

Total_Payload_Mass
45596.0

- Calculate the total payload carried by boosters from NASA
- Query in screenshot, the results displays the total payload mass.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) AS Avg_Payload_Mass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

Avg_Payload_Mass
2928.4

- Calculate the average payload mass carried by booster version F9 v1.1
- Query in screenshot, the result displays the average payload mass carried by booster version F9 v 1.1

First Successful Ground Landing Date

```
%%sql
SELECT MIN("Date") AS First_Successful_Landing
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

First_Successful_Landing
2015-12-22

- Find the dates of the first successful landing outcome on ground pad
- Query in screenshot, the result shows the date of the first successful landing outcome on a ground pad: 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
  AND "Payload_Mass__kg_" > 4000
  AND "Payload_Mass__kg_" < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Query in screenshot, the result lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Query in screenshot, the result shows the number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Outcome_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Payload_Mass__KG_" = (
    SELECT MAX("Payload_Mass__KG_")
    FROM SPACEXTABLE
);
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- Query in screenshot, the result lists the names of the booster which have carried the maximum payload mass.

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
%%sql
SELECT
    substr("Date", 6, 2) AS Month,
    "Booster_Version",
    "Launch_Site",
    "Landing_Outcome"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND substr("Date", 1, 4) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Query in screenshot, the result lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY Outcome_Count DESC;
```

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

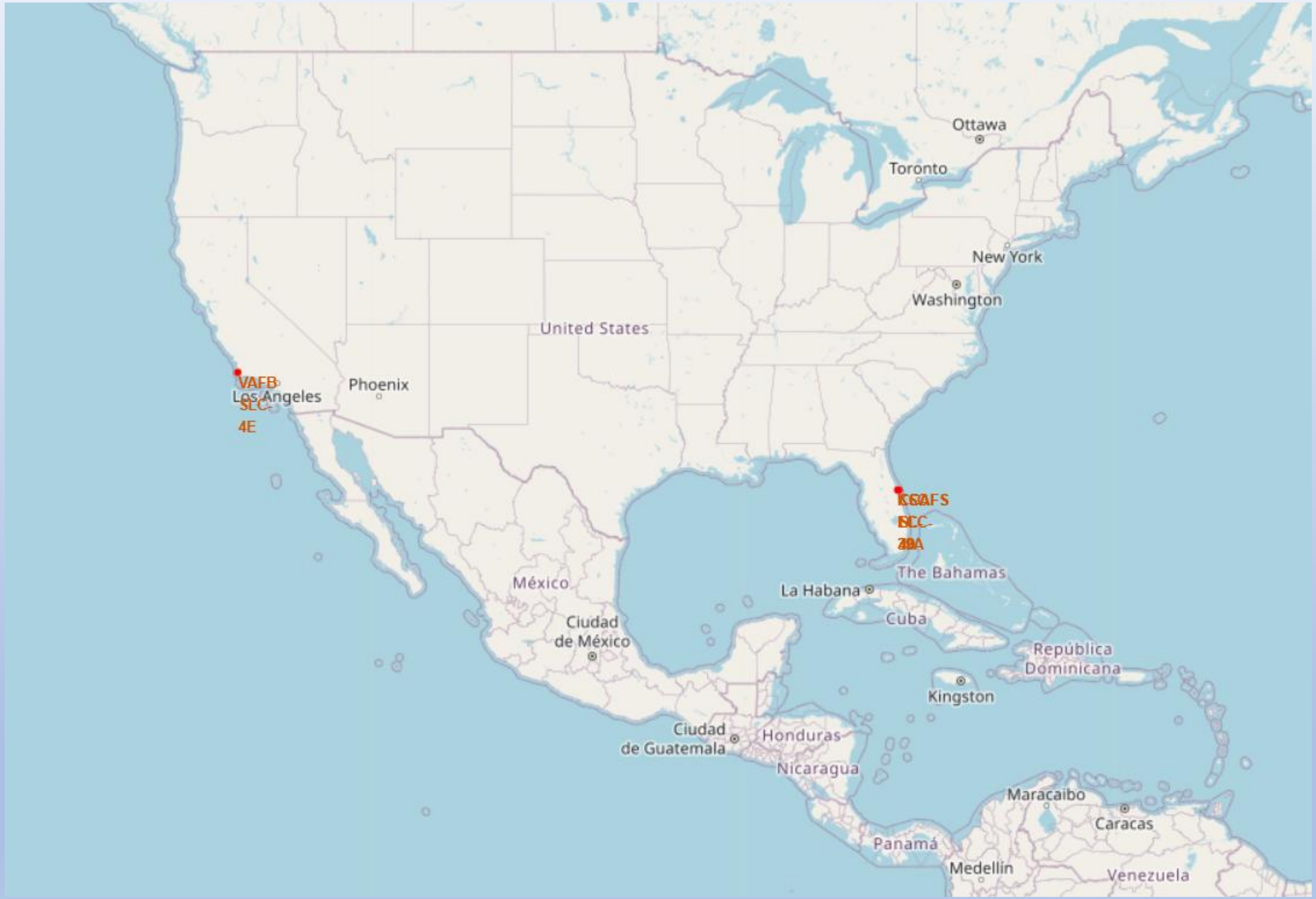
- Query in screenshot, the result displays the count rank of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

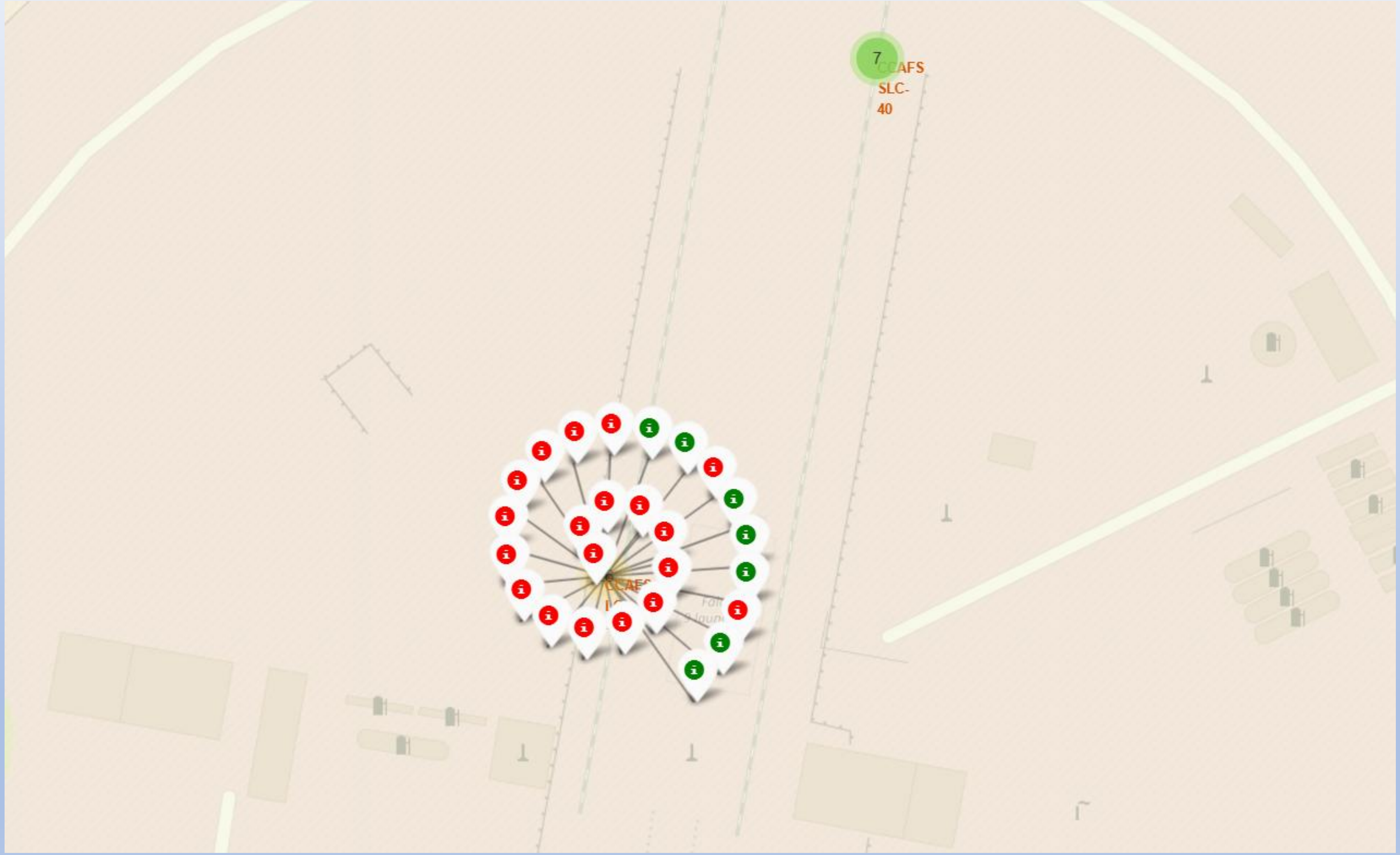
Map of all Launch Sites

- All launch sites are in close proximity to coastlines and in areas with mild weather.



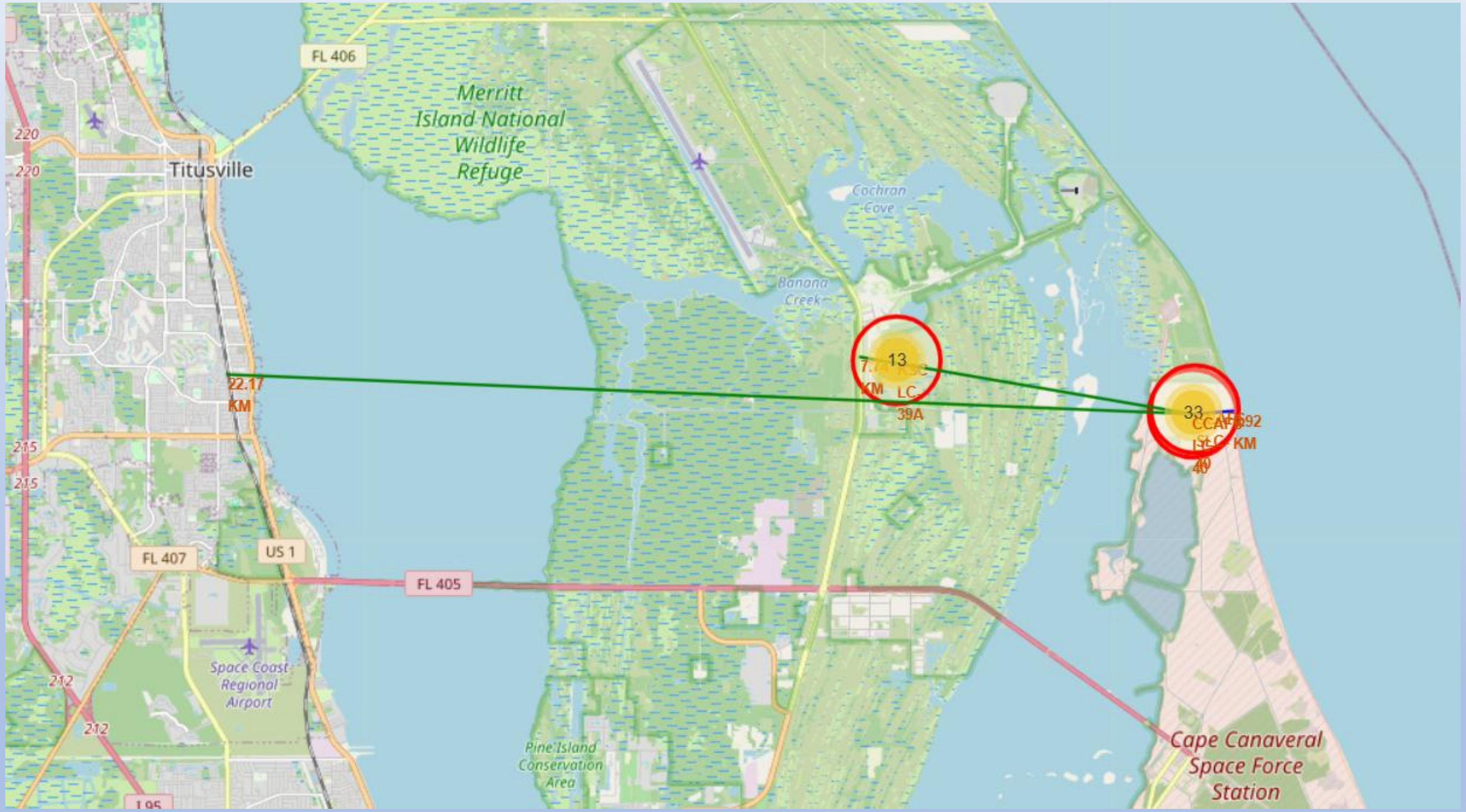
Mark the success/failed launches for each site on the map

- For launch site CCAFS LC-40 the Green markers indicate a successful launch, while the Red markers indicate an unsuccessful launch. Similar data will be displayed when exploring the interactive map.



Map of distances between a launch site and its proximities

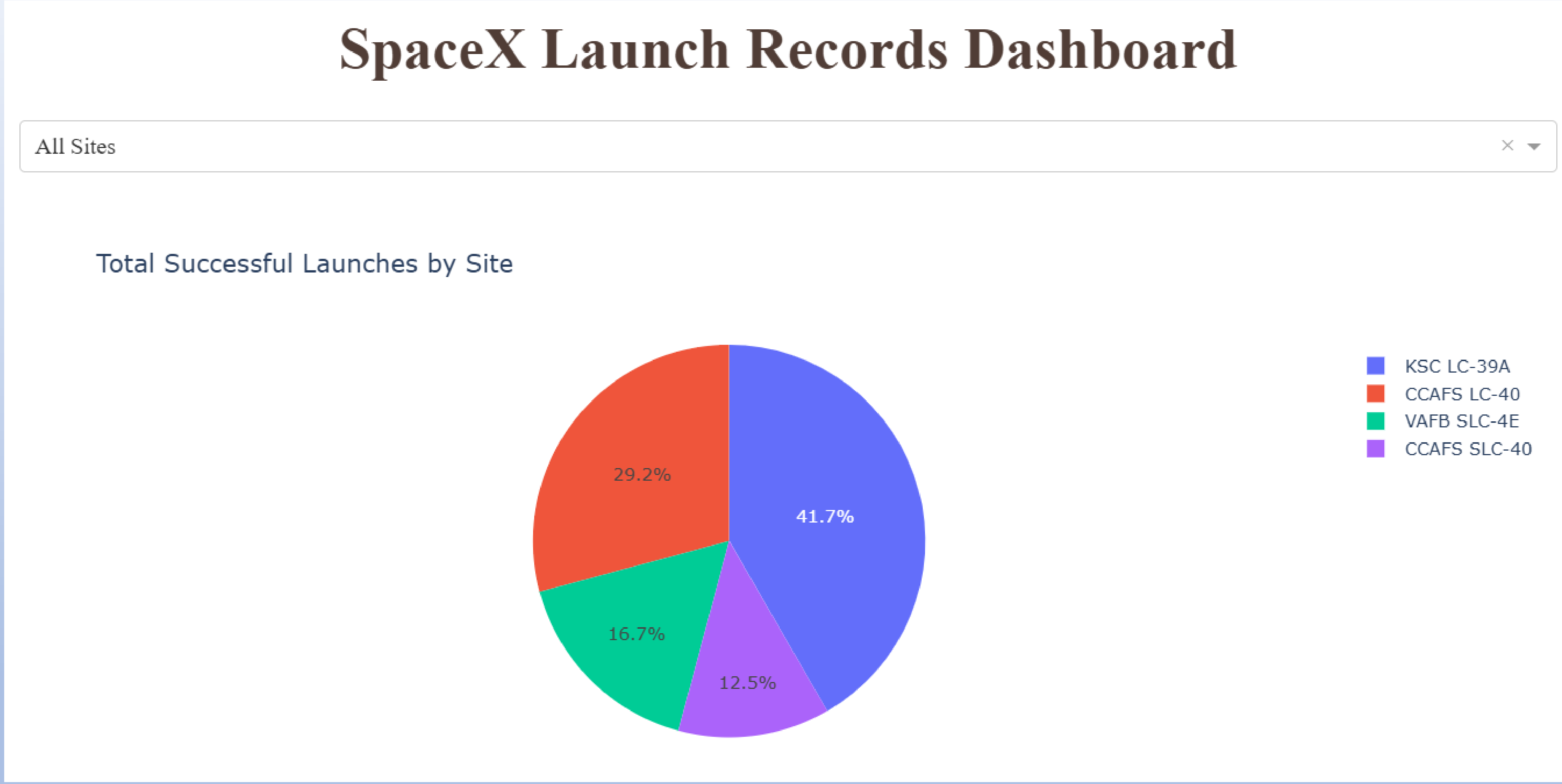
- Interactive map that calculates the distance from launch site CCAFS LC-40 to it proximities, such as the coastline, nearest highway and railroad.



Section 4

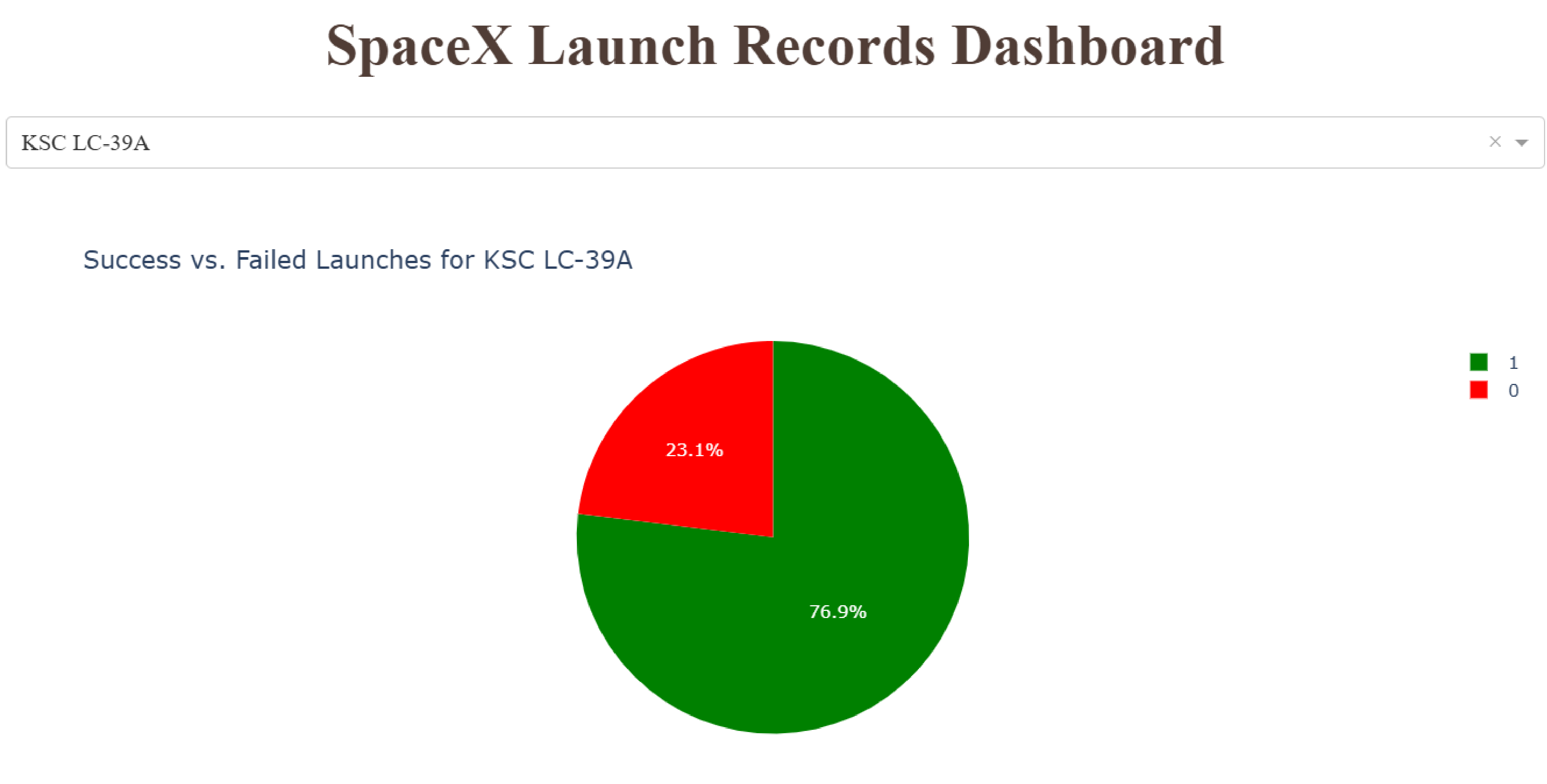
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard – Total Successful Launches by Site



- Interactive Pie Chart with dropdown menu. Current screen shot shows the Total Successful Launches by Site. KSC LC-39A is the most successful.

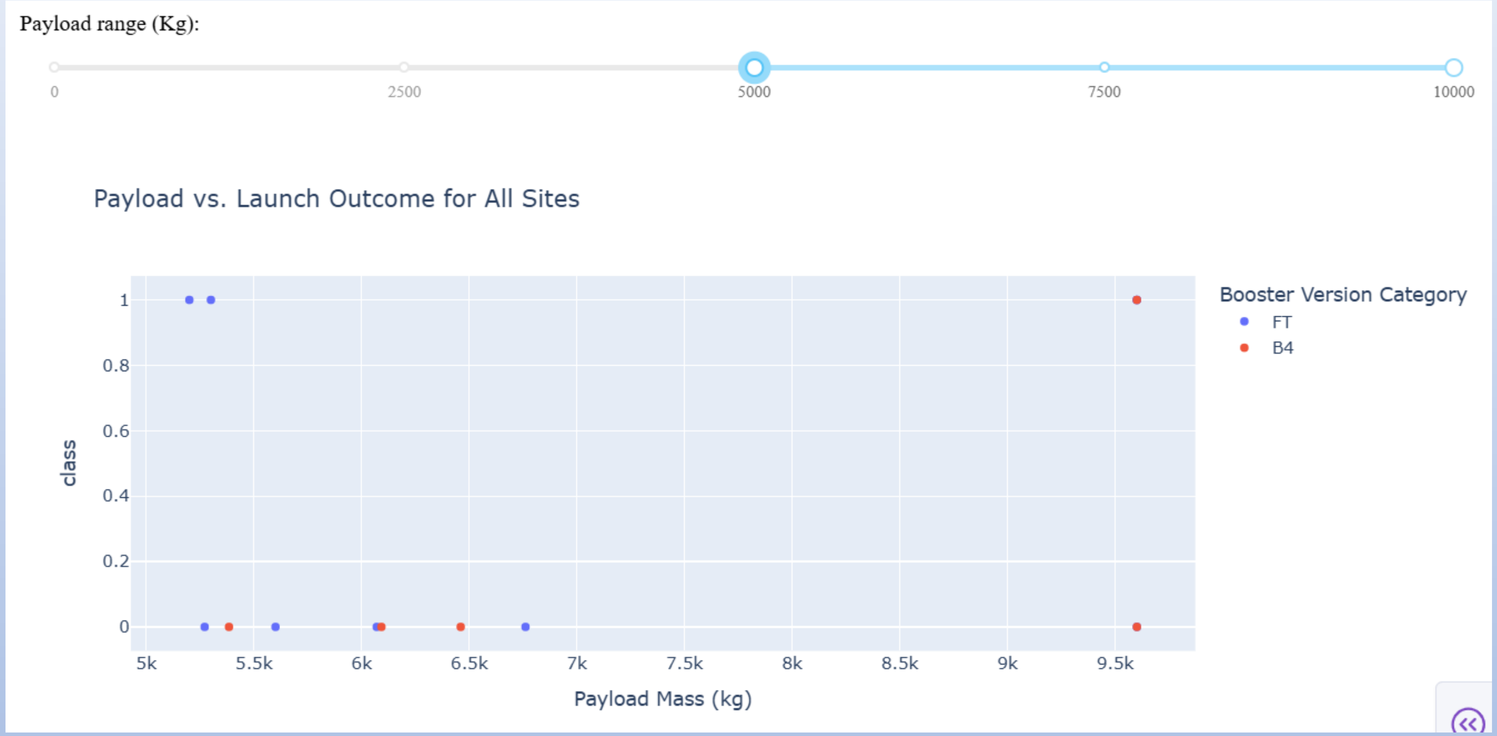
Success vs. Failed Launches for KSC LC-39A



- Interactive dashboard of a Pie chart. Current display shows the high launch success rate at site KSC LC-39A.

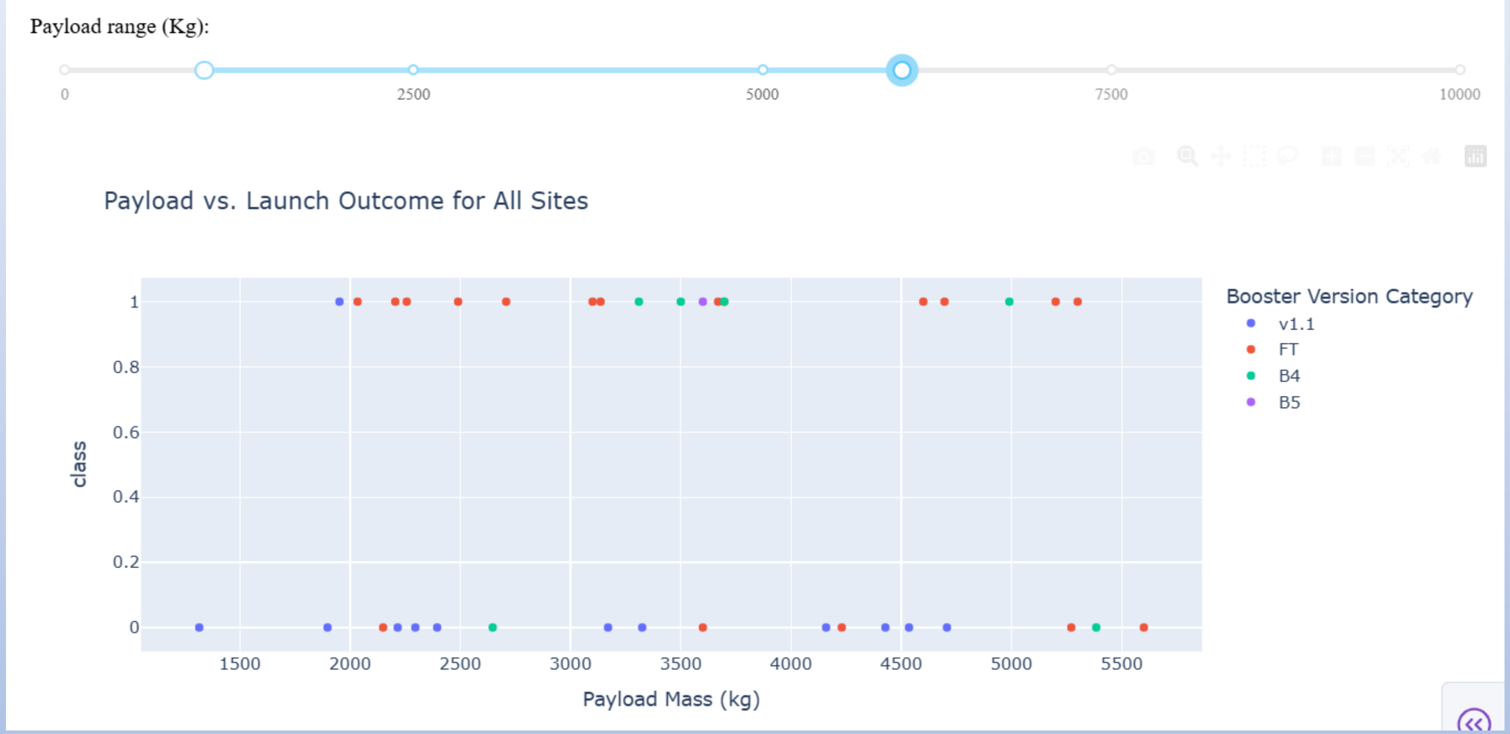
Payload vs. Launch Outcome for All Sites

- For payloads from 5,000 Kg to 10,000 Kg only two types of boosters were used, FT and B4. FT has a higher success rate in the payload range.



Payload vs. Launch Outcome for All Sites

- For payloads from 1,000 Kg to 6,000 Kg all booster types were used. FT appears to have the highest success rate in this payload range.

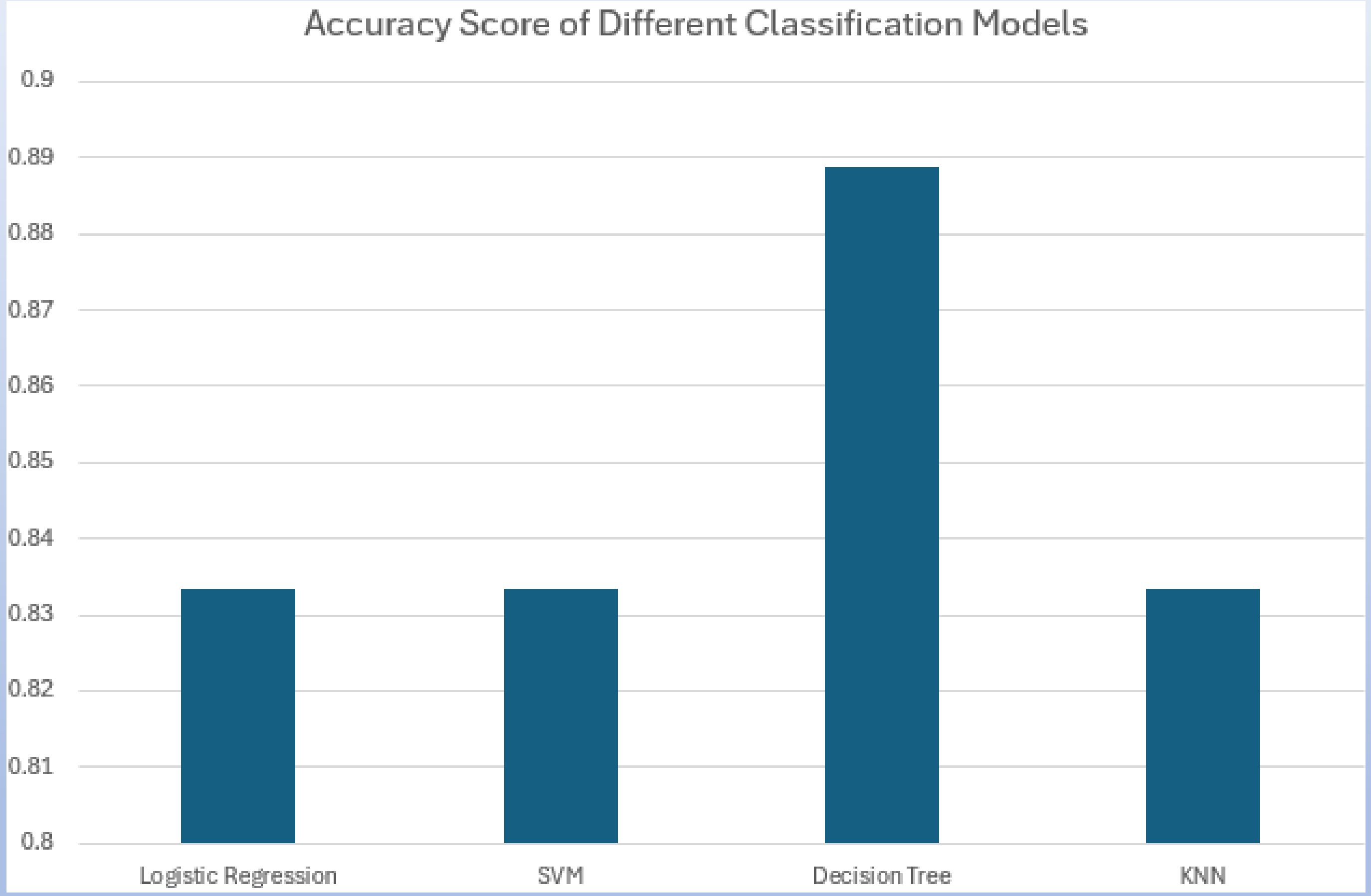


Section 5

Predictive Analysis (Classification)

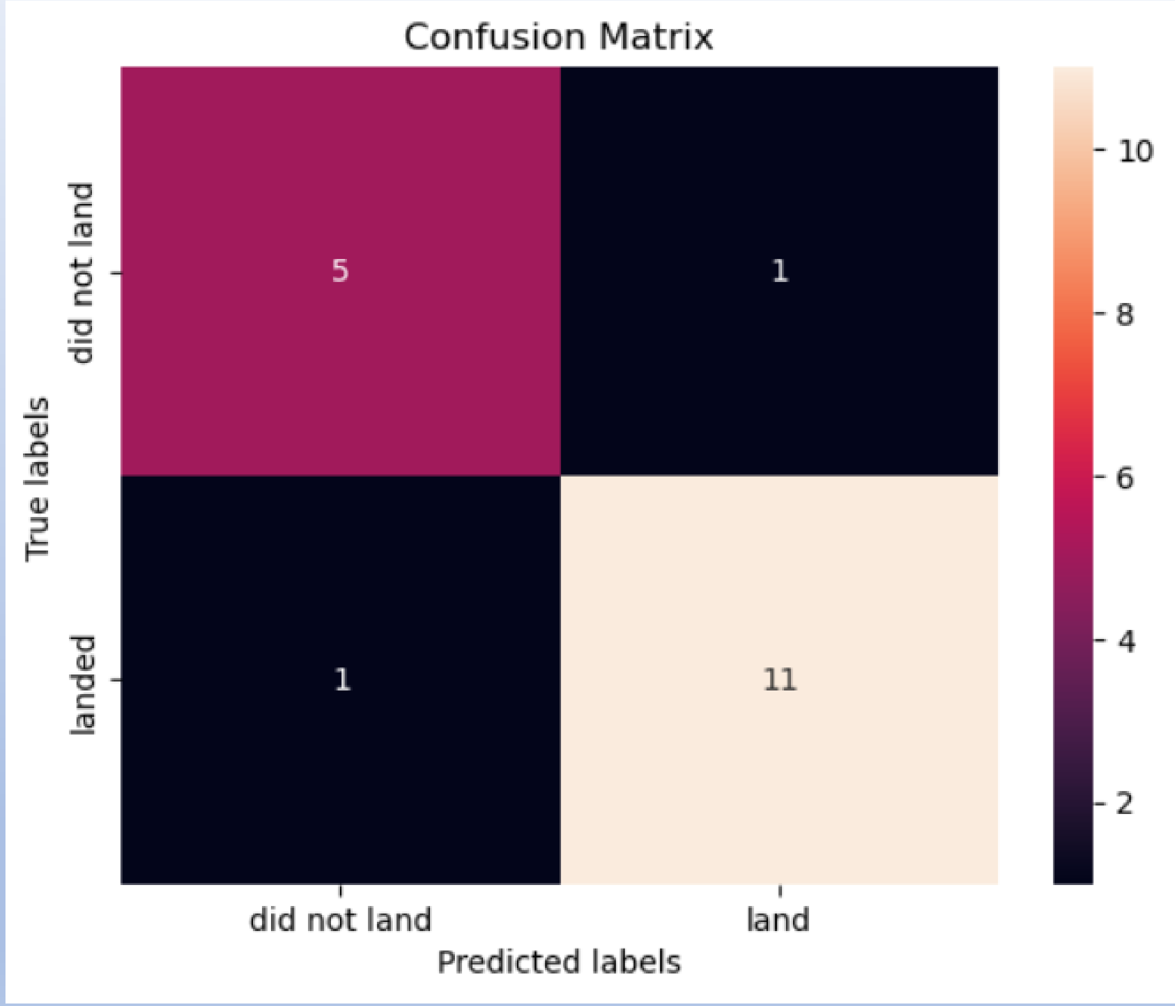
Classification Accuracy

- The Decision Tree model has the highest accuracy score of 0.88888888. Logistic Regression, SVM and KNN all scored 0.8333333.



Confusion Matrix

- Confusion Matrix on the Decision Tree model shows only 1 false negative and 1 false positive.



Conclusions

• The Decision Tree Classifier emerged as the most effective machine learning model for predicting Falcon 9 booster landing success.

• Payload mass plays a significant role in landing outcomes, with lighter payloads achieving higher success rates than heavier ones.

• Launch success has improved consistently over time, indicating that SpaceX's landing reliability increases with experience and operational refinement.

• Among all launch sites, KSC LC-39A demonstrated the highest success rate, highlighting its operational effectiveness.

• Missions to GEO, HEO, SSO, and ES-L1 orbits showed the strongest landing success rates.

• Analysis of booster versions revealed that the "FT" booster achieved high success rates across a wide range of payload masses, demonstrating strong reliability and robustness.

Appendix

- GitHub repository URL: Applied Data Science Capstone Project Repository Hyperlink

Thank you!